

Whole-genome annotation at NCBI

François Thibaud-Nissen (thibaudf@ncbi.nlm.nih.gov)

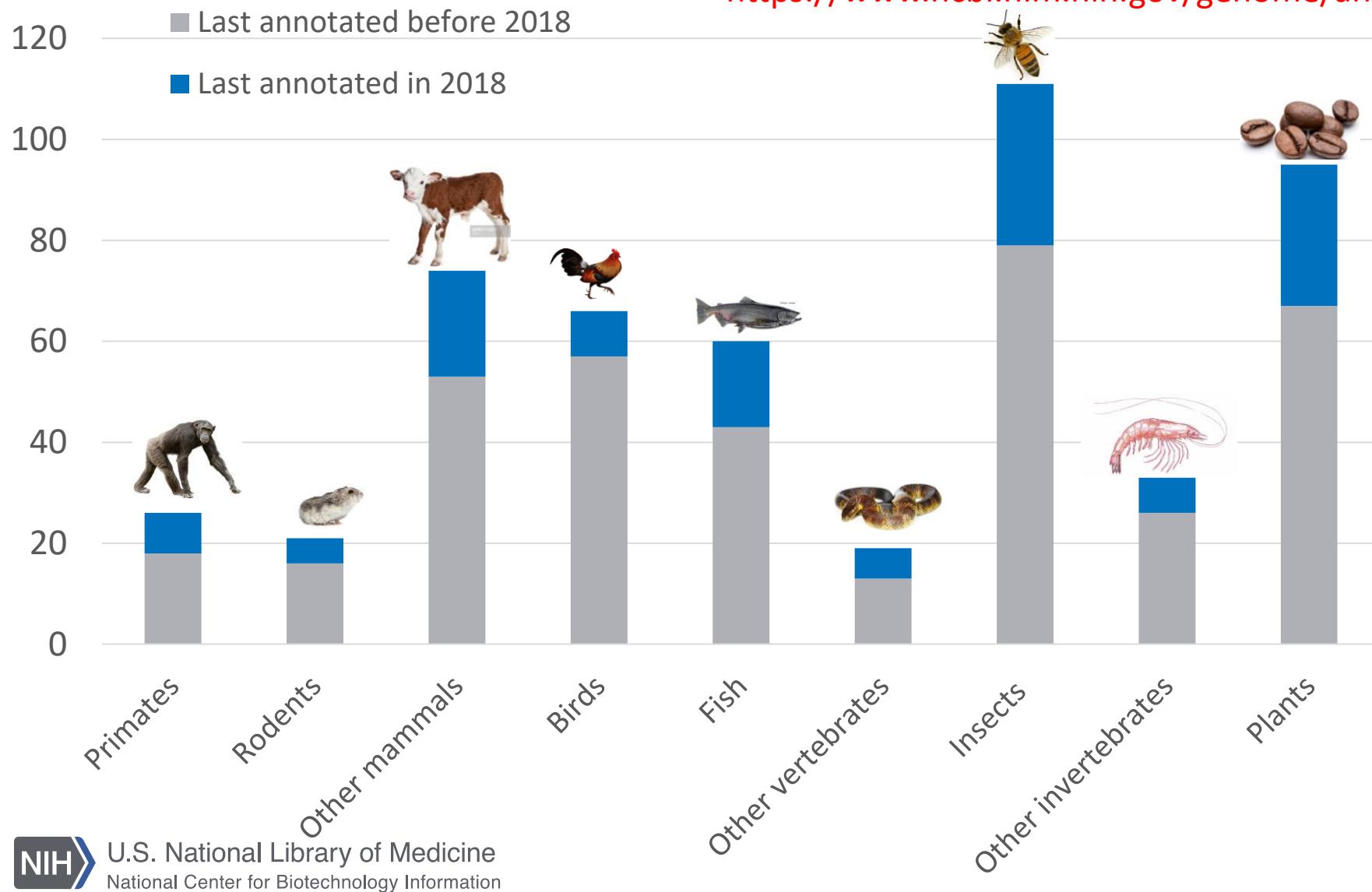
PAG XXVII January 13, 2019



U.S. National Library of Medicine
National Center for Biotechnology Information

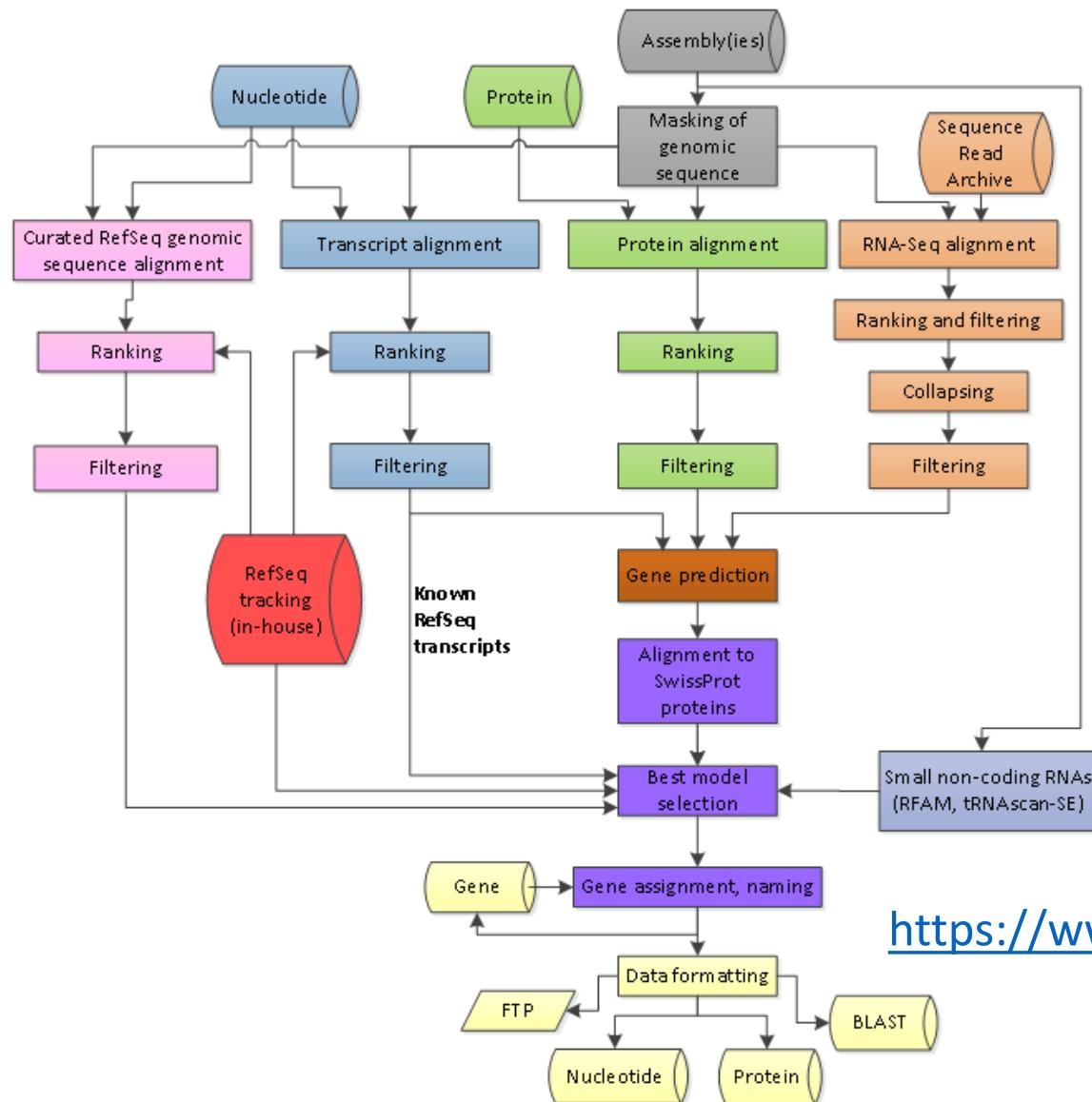
Over 500 eukaryotic species annotated

https://www.ncbi.nlm.nih.gov/genome/annotation_euk/all/



U.S. National Library of Medicine
National Center for Biotechnology Information

The NCBI Eukaryotic Genome Annotation Pipeline



- Automated
- 1 to 2 weeks start to end
- Consumes public data!
- Evidence driven:
 - Same species and close cross-species transcripts, proteins
 - RNA-Seq and IsoSeq
 - TSA (Transcript Shotgun Assemblies)

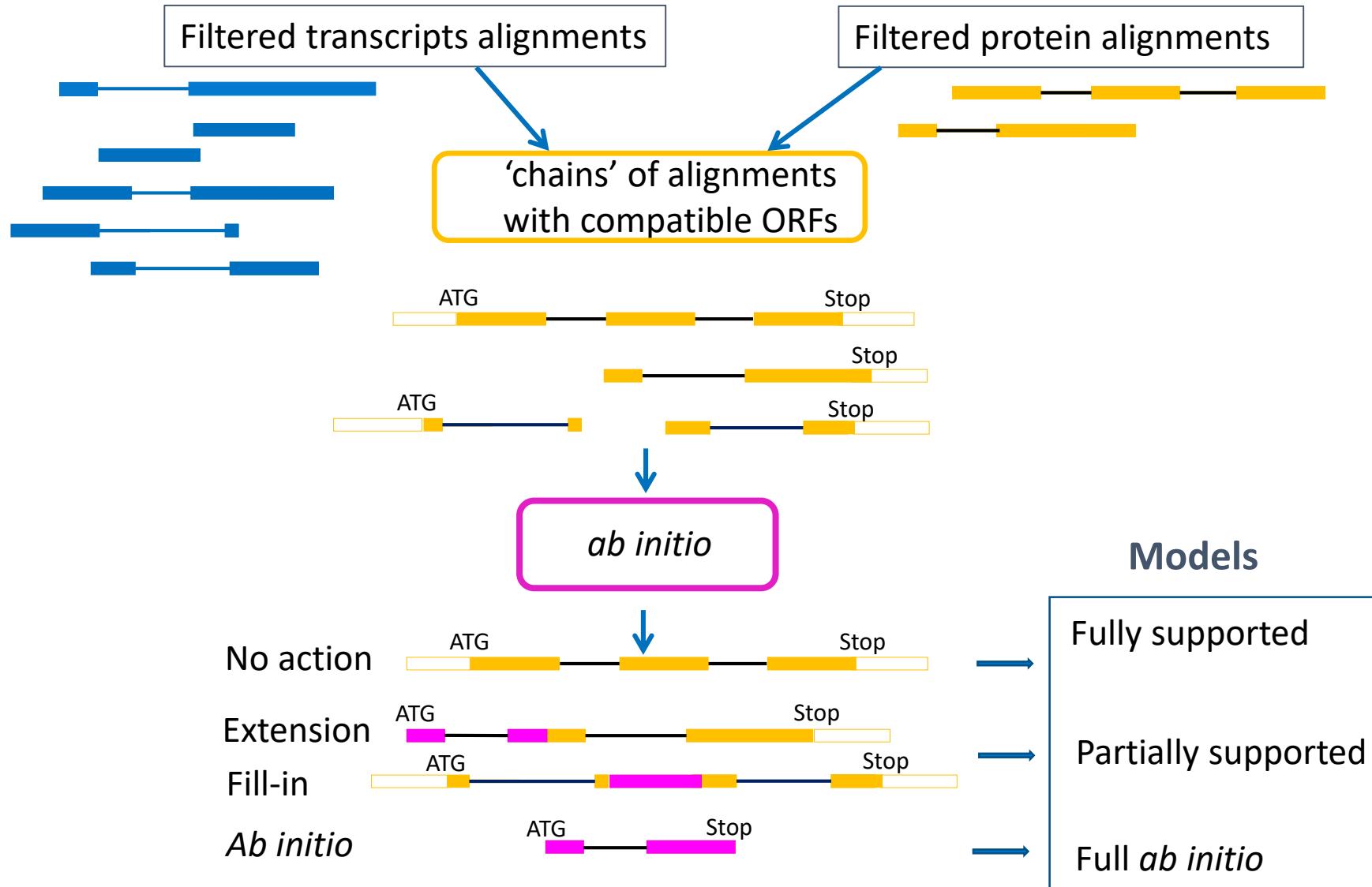
https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/



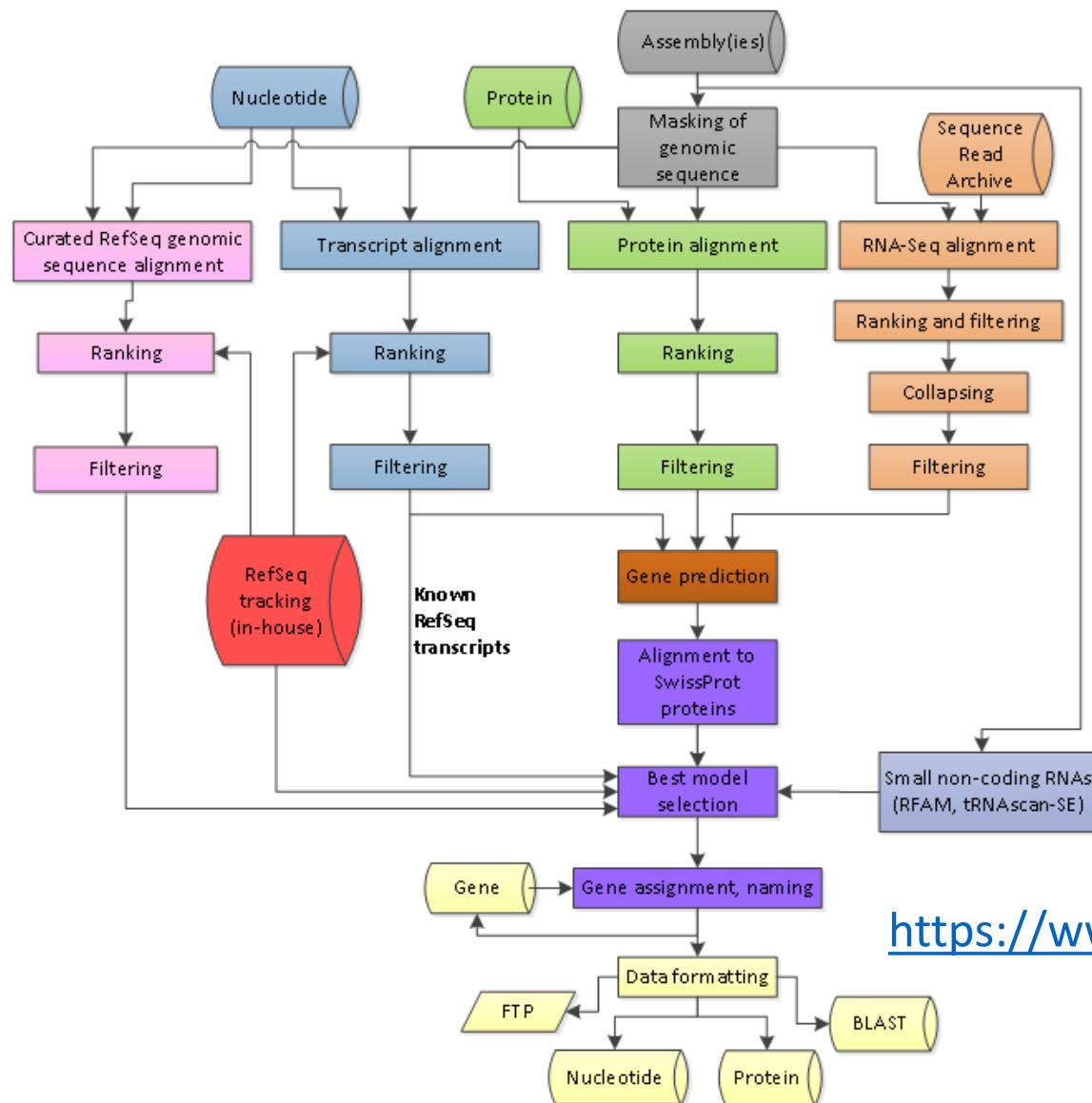
U.S. National Library of Medicine
National Center for Biotechnology Information



Gnomon gene finder



The NCBI Eukaryotic Genome Annotation Pipeline



- Automated
- 1 to 2 weeks start to end
- Consumes public data!
- Evidence driven:
 - Same-species and close cross-species transcripts, proteins
 - RNA-Seq and IsoSeq
 - TSA (Transcript Shotgun Assemblies)

https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/



U.S. National Library of Medicine
National Center for Biotechnology Information



RefSeq nomenclature for eukaryotes

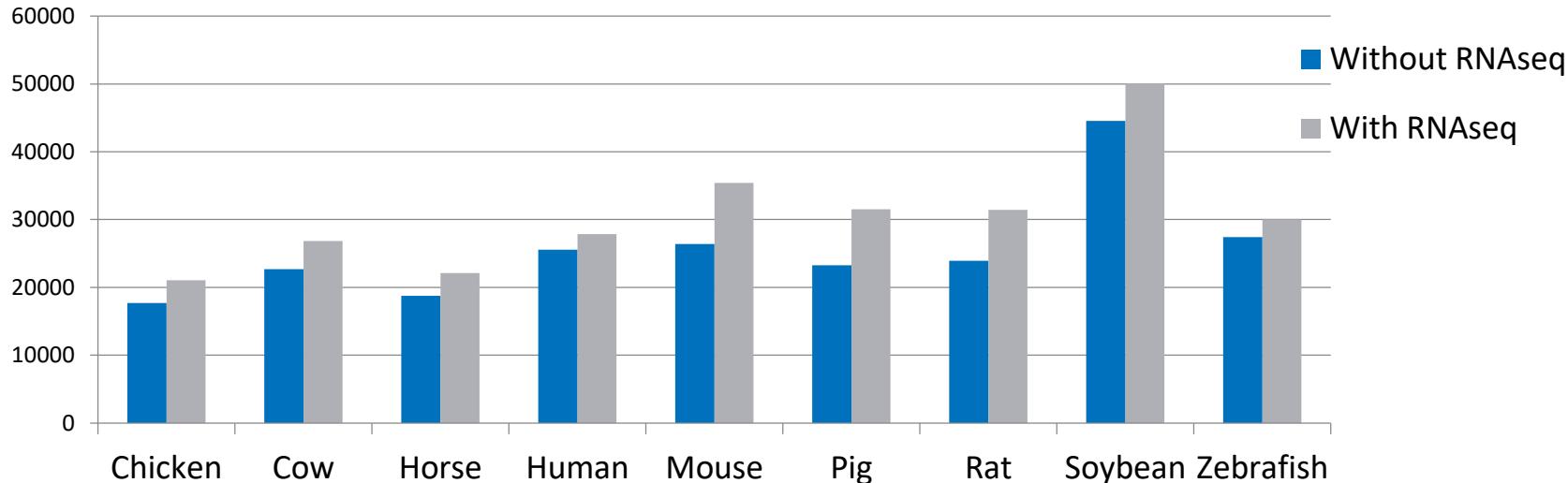
Model RefSeq = XM_*, XR_* or XP_*	Known RefSeq = NM_*, NR_* or NP_* prefixes
<ul style="list-style-type: none">Produced by the annotation pipelineThe vast majority is fully supported by experimental evidence	<ul style="list-style-type: none">Maintained by the RefSeq curation staffCurated for only a small number of organisms<ul style="list-style-type: none">HumanMouseCorn, rat, soybean, chicken, tomato, cow, pig, xenopus

The annotation for a genome comprises BOTH sets of transcripts
Model RefSeq should not be excluded from analysis

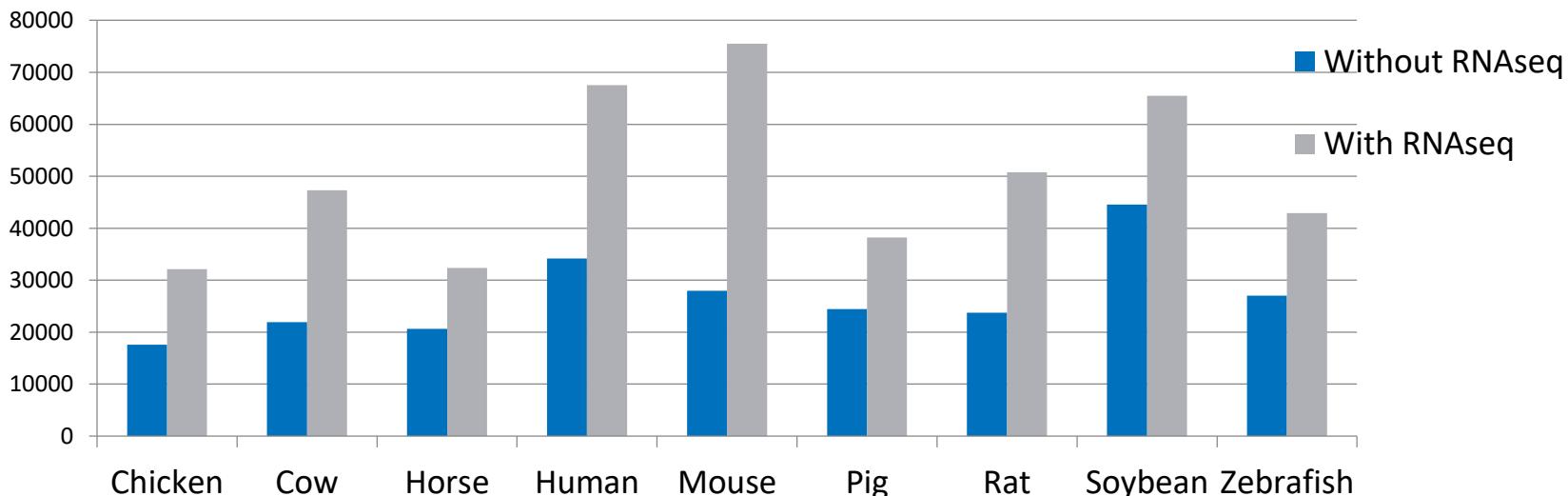


Heavy use of RNA-Seq data

Number of genes predicted +/- RNaseq



Number of coding transcripts predicted +/- RNaseq

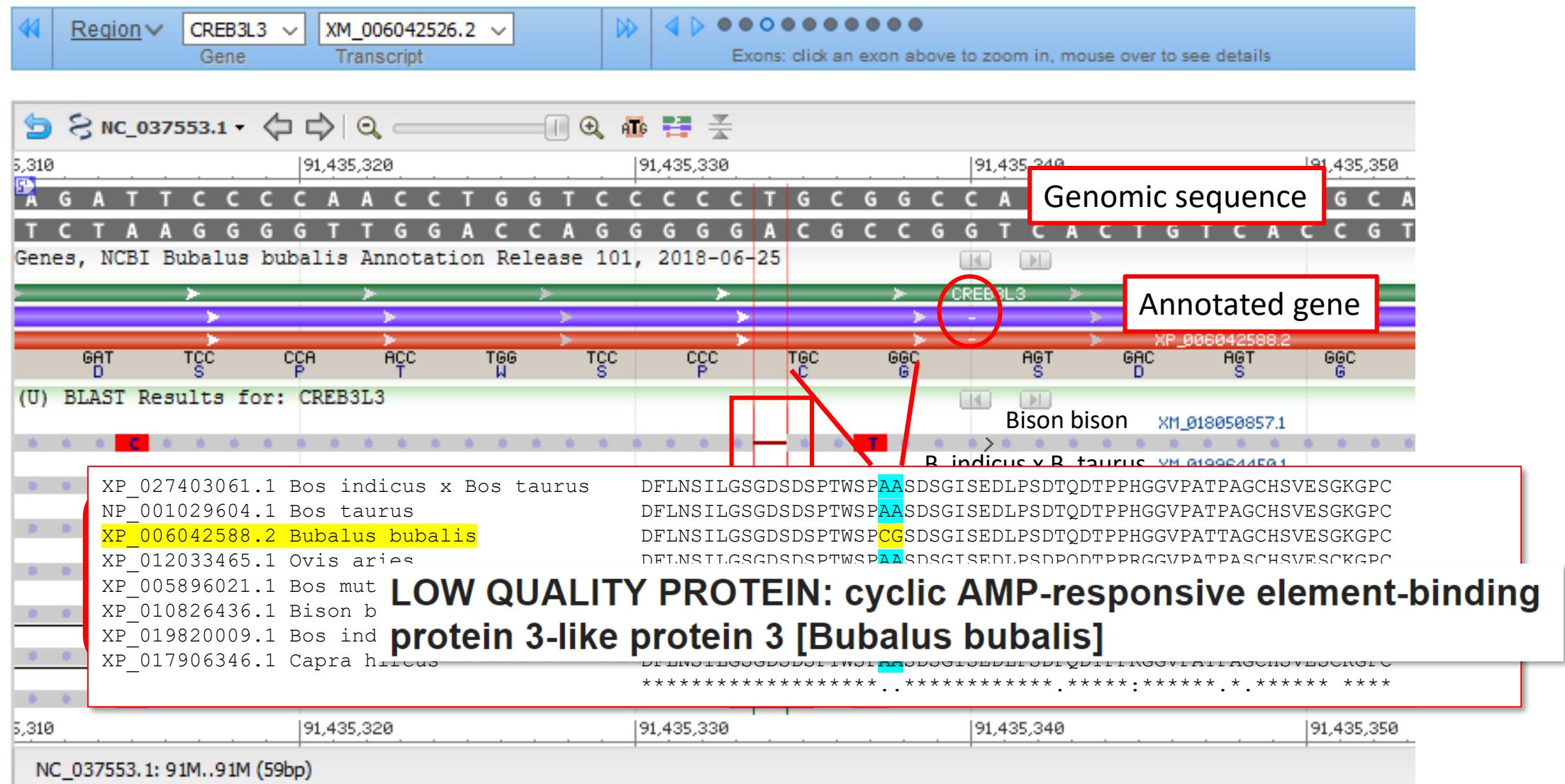


Quality metrics for the evaluation of annotation and assemblies

Number of	Too high	Too low
genes	<ul style="list-style-type: none">• Allelic duplication• Contamination	<ul style="list-style-type: none">• Missing sequence in assembly• Too little evidence
genes with partial support	<ul style="list-style-type: none">• Too little evidence• Fragmented assembly• Distant species	✓
genes with no support	<ul style="list-style-type: none">• Contamination• Poor masking (TEs)• Too little evidence• Distant species	✓
genes with orthologs	✓	<ul style="list-style-type: none">• Contamination• Allelic duplication
partial CDS	<ul style="list-style-type: none">• Fragmented assembly• Bad order and orientation	✓
'corrected' CDSs	High error rate of the assembly sequence	✓



Deletion in gene model to compensate for error in genome: CREB3L3

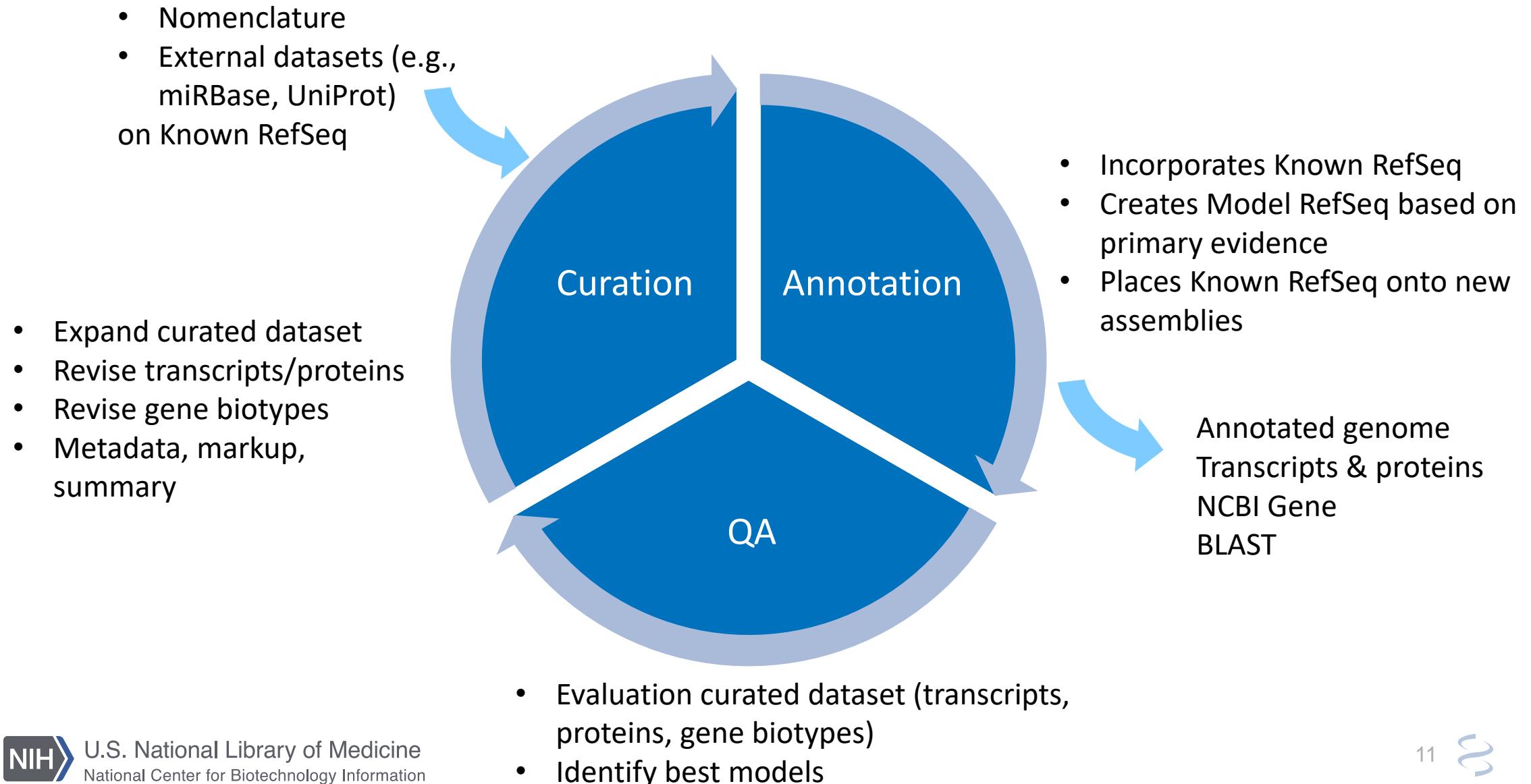


High percentage of corrected genes in poorly polished PacBio assemblies

Organism name	Assembly name	Assembly accession	Submission date	Sequencing technology	Percent corrected coding genes
<i>Myotis lucifugus</i>	Myoluc2.0	GCF_000147115.1	14-Sep-10	Sanger	12
<i>Myotis davidii</i>	ASM32734v1	GCF_000327345.1	20-Dec-12	Illumina HighSeq 2000	8
<i>Myotis brandtii</i>	ASM41265v1	GCF_000412655.1	28-Jun-13	Illumina HiSeq 2000	5
<i>Pteropus alecto</i>	ASM32557v1	GCF_000325575.1	4-Dec-13	Illumina HighSeq 2000	6
<i>Pteropus vampyrus</i>	Pvam_2.0	GCF_000151845.1	5-Dec-14	Illumina	4
<i>Rousettus aegyptiacus</i>	ASM146680v1	GCF_001466805.1	15-Dec-15	Illumina HiSeq; PacBio	57
<i>Rousettus aegyptiacus</i>	Raegyp2.0	GCF_001466805.2	17-Mar-16	Illumina HiSeq; PacBio	11
<i>Miniopterus natalensis</i>	Mnat.v1	GCF_001595765.1	22-Mar-16	Illumina HiSeq	4
<i>Rhinolophus sinicus</i>	ASM188883v1	GCF_001888835.1	4-Dec-16	Illumina HiSeq	4
<i>Hipposideros armiger</i>	ASM189008v1	GCA_001890085.1	6-Dec-16	Illumina HiSeq	3
<i>Desmodus rotundus</i>	ASM294091v2	GCF_002940915.1	23-Feb-18	Illumina Hiseq 2000/2500	6

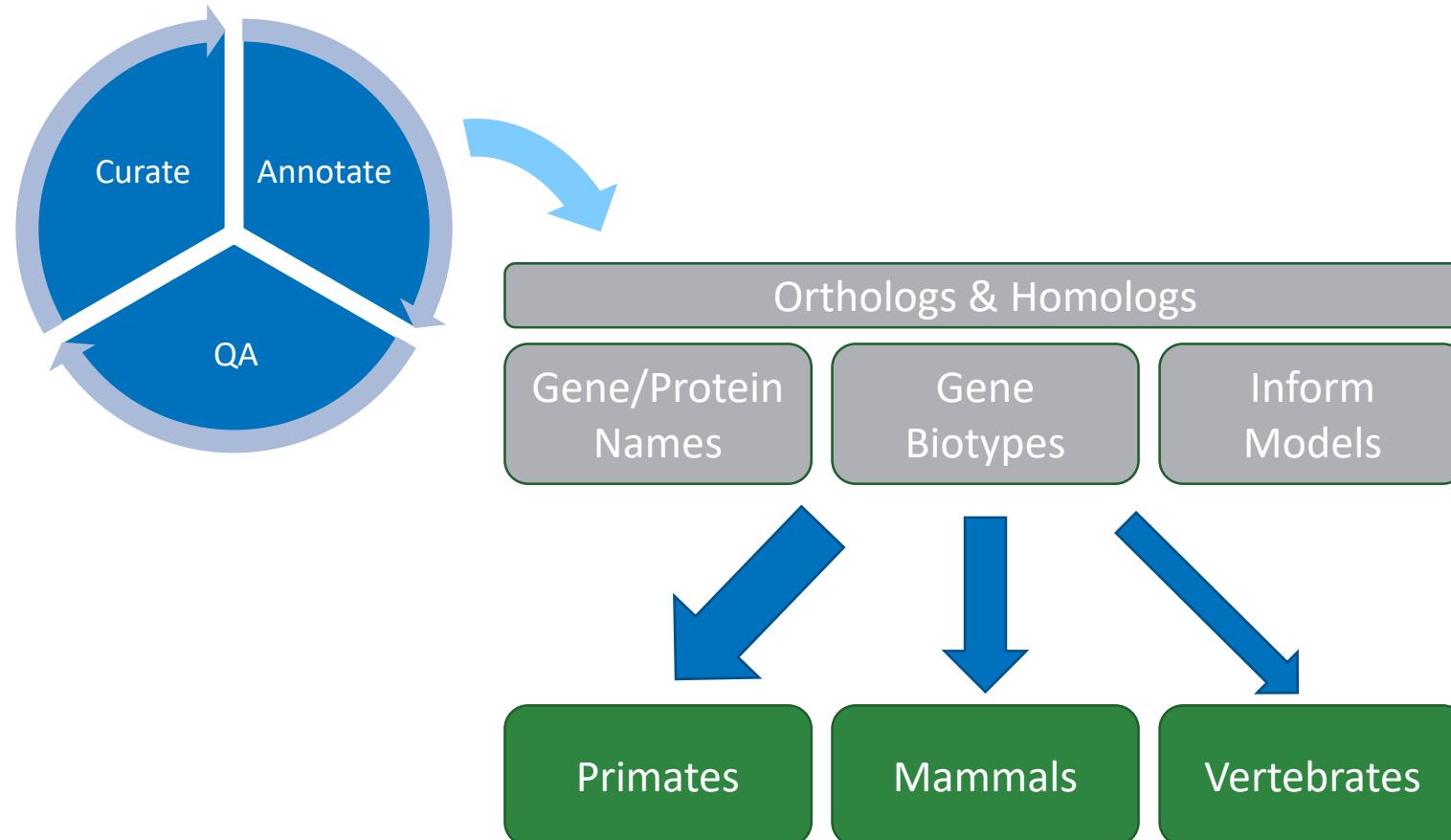


Interplay of curation and annotation



How curated data informs annotation

Reference species curation (e.g. human)



Where is the annotation?

https://www.ncbi.nlm.nih.gov/genome/annotation_euk/all/

▼ Fish (60)

FTP - FTP Download **B** - Organism-specific BLAST **AR** - Annotation Report **GDV** - Genome Data Viewer

Species	RefSeq assembly(ies)	Annotation Release	Freeze Date	Release Date	Links
Larimichthys crocea (large yellow croaker)	L_crocea_2.0 (GCF_000972845.2)	102	2018-12-04	2018-12-05	FTP B AR GDV
Tachysurus fulvidraco (yellow catfish)	ASM372403v1 (GCF_003724035.1)	100	2018-11-19	2018-11-26	FTP B AR GDV
Electrophorus electricus (electric eel)	Ee_SOAP_WITH_SSSPACE (GCF_003665695.1)	100	2018-11-06	2018-11-08	FTP B AR GDV



Up to higher level directory

Name

- ARCHIVE
- Assembled_chromosomes
- CHR_A1
- CHR_A2
- CHR_A3

Download

blastn blastp blastx tblastn tblastx

Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
[Clear](#) [Query subrange](#)

From To

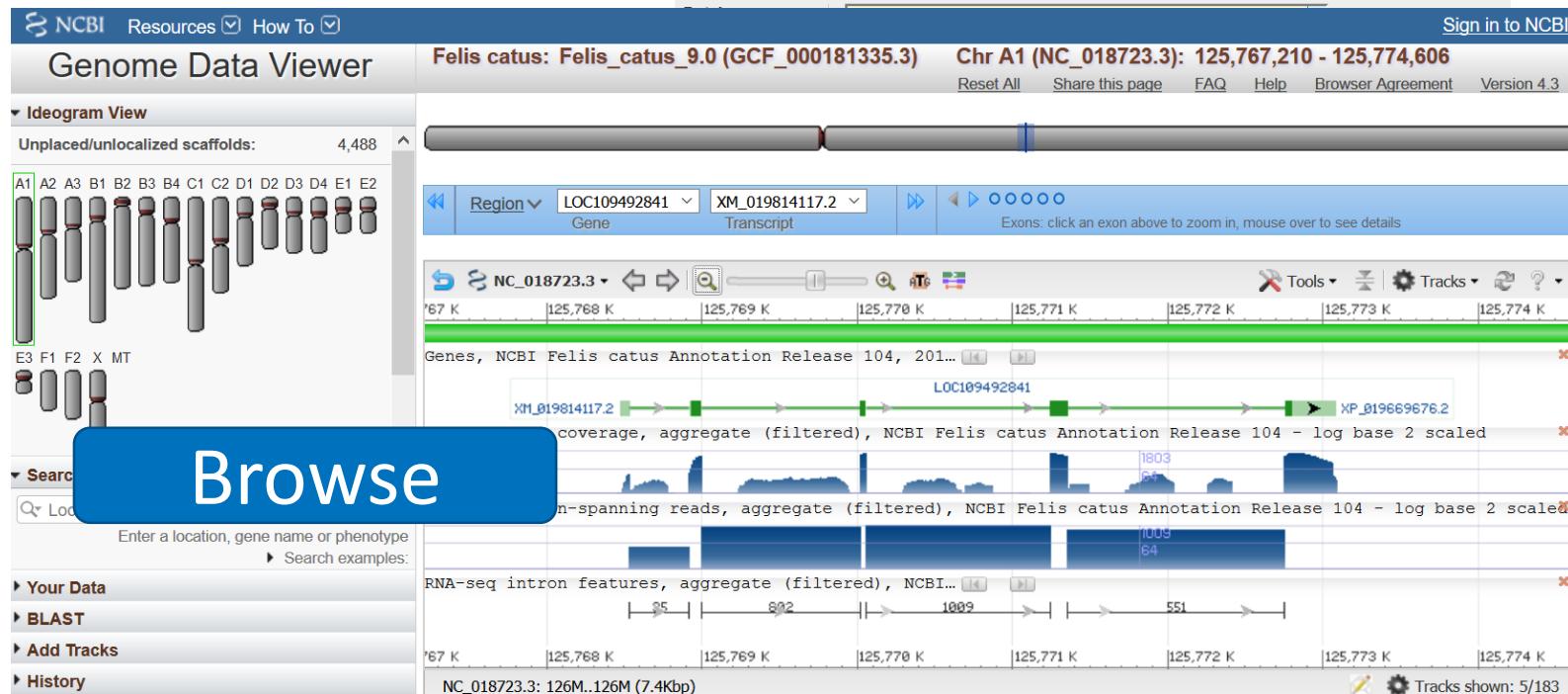
BLAST

Or, upload file No file selected.

Job Title

Enter a descriptive title for your BLAST search

Choose Search Set



Summary report

Annotation Release information

This annotation should be referred to as NCBI [Felis catus](#) Annotation Release 104.

Annotation release ID: 104

Date of Entrez queries for transcripts and proteins: Dec 6 2017

Date of submission of annotation to the public databases: Dec 6 2017

Software version: [8.0](#)

Assemblies

The following assemblies were included in this annotation run:

Assembly name	Assembly accession	Submitter
Felis catus 9.0	GCF_000181335.3	Genome Sequencing Center
Felis_catus_9.0 (Current) to Felis_catus_8.0 (Previous)		
Identical		19%
Minor changes		49%
Major changes		16%
New		17%
Deprecated		17%
Other		<1%
Download the report		tabular , Genome Workbench

Gene and feature statistics

Counts and length of annotated features are provided below for each assembly.

Feature counts

Feature	Felis_catus_9.0
Genes and pseudogenes	35,588
protein-coding	19,748
non-coding	11,669
transcribed pseudogenes	3
non-transcribed pseudogenes	4,082
genes with variants	12,732
immunoglobulin/T-cell receptor gene segments	86
other	0
rRNAs	54,713
fully-supported	53,394
with > 5% ab initio	613
partial	182
with filled gap(s)	0
known RefSeq (NM_)	378
model RefSeq (XM_)	54,335
non-coding RNAs	18,016

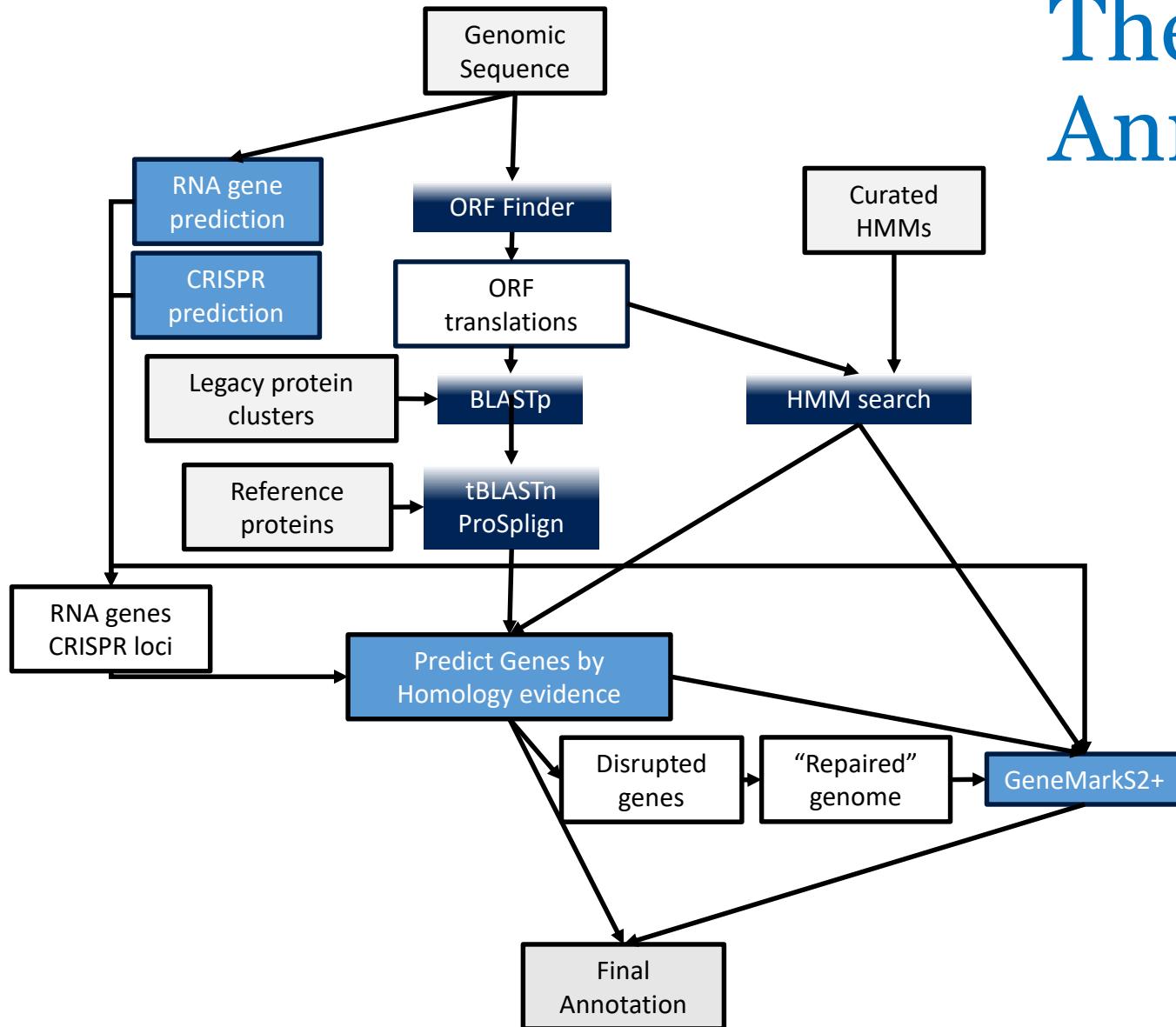


To sum up...

- Driven by primary evidence, RNAseq in particular
- Depends on assembly quality
- Benefits from RefSeq curation (Vertebrates, plants)
- Publicly available



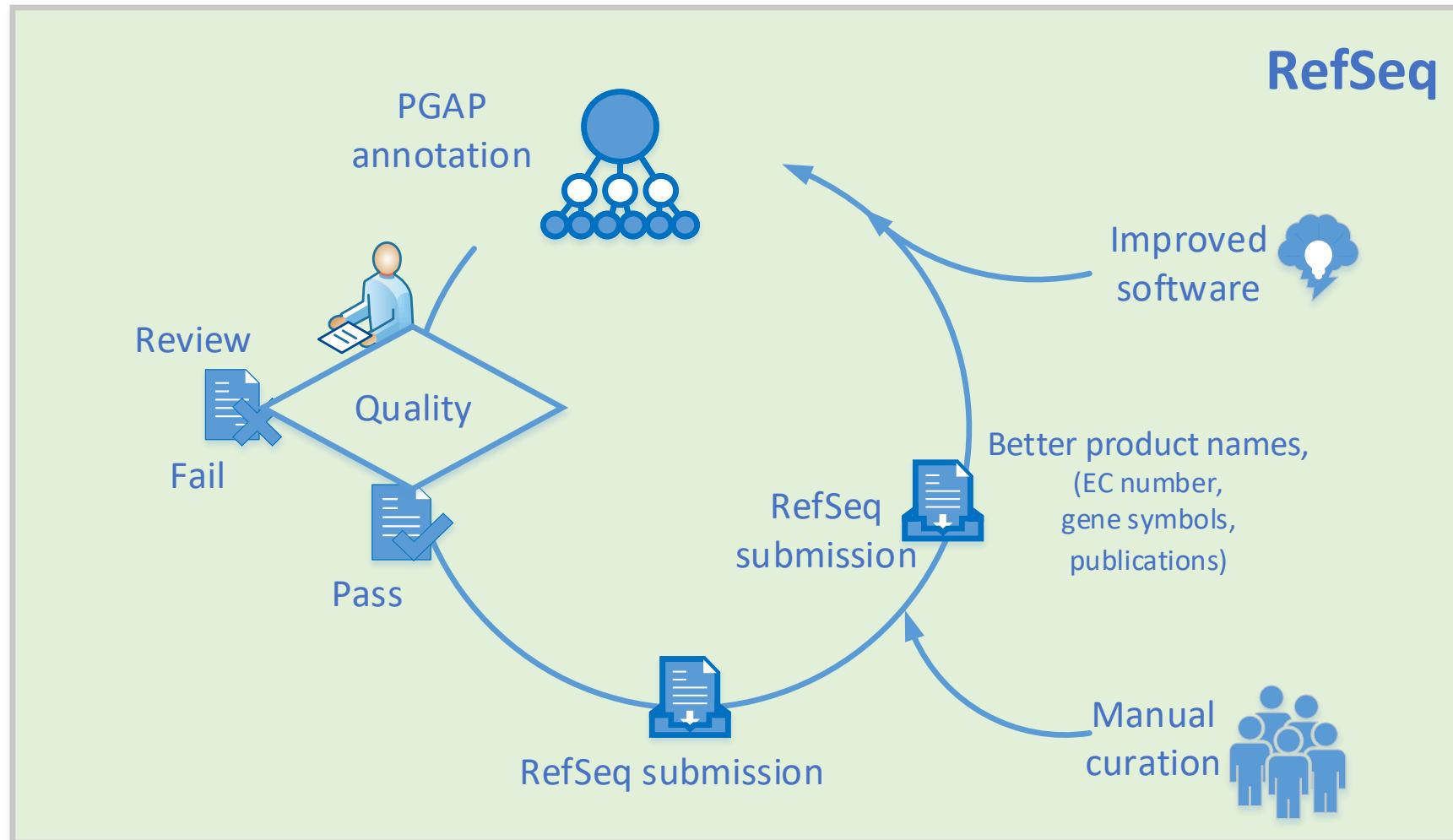
The Prokaryotic Genome Annotation Pipeline



- Automated
- Protein-coding genes prediction
 - Protein homology
 - Ab initio calls (GeneMarkS2+)
 - Hidden Markov Models
- Non-coding genes
 - tRNA
 - CRISPER loci
 - rRNAs
- Functional annotation
 - Hidden Markov Models
 - BlastRules
 - CDD architectures
 - Protein homology



Continuous updates of RefSeq annotation



Improvements of functional annotation

- Addition or improvement of protein product names, gene symbols, EC numbers based on the literature
- Approach: curate of ‘annotation rules’ that support proteins, rather proteins themselves
 - Hidden Markov Models
 - TIGRFAMs
 - PFAMs
 - HMMs for antimicrobial resistance genes
 - HMMs based on NCBI protein clusters
 - BlastRules – one or more protein + identity cutoff + coverage cutoff
 - Transposable elements
 - CDD architectures



Non redundant protein model

WP_001102383.1: single protein annotated on 4019 distinct genome assemblies

GenPept ▾

 This record is a non-redundant protein sequence. Please [read more here](#).

MULTISPECIES: acidic protein MsyB [Enterobacteriaceae]

NCBI Reference Sequence: WP_001102383.1

[Identical Proteins](#) [FASTA](#) [Graphics](#)

Go to:

LOCUS
DEFINITION
ACCESSION
VERSION
KEYWORDS
SOURCE
ORGANISM

COMMENT

	Source	CDS Region in Nucleotide	Protein	Name	Organism	Strain	Assembly
	RefSeq	NC_009800.1 10642-11355 (-)	WP_001102383.1	acidic protein MsyB	Escherichia coli HS	HS	GCF_000017765.1
	RefSeq	NC_009801.1 11921-12634 (-)	WP_001102383.1	acidic protein MsyB	Escherichia coli O139:H28 str. E24377A	E24377A	GCF_000017745.1
	RefSeq	NC_013361.1 10643-11356 (-)	WP_001102383.1	acidic protein MsyB	Escherichia coli O26:H11 str. 11368	11368	GCF_000091005.1
	RefSeq	NC_013364.1 10643-11356 (-)	WP_001102383.1	acidic protein MsyB	Escherichia coli O111:H- str. 11128	11128	GCF_000010765.1
	RefSeq	NC_016902.1 4099268-4099981 (+)	WP_001102383.1	acidic protein MsyB	Escherichia coli K011FL	KO11	GCF_000147855.2

Taxonomic Groups

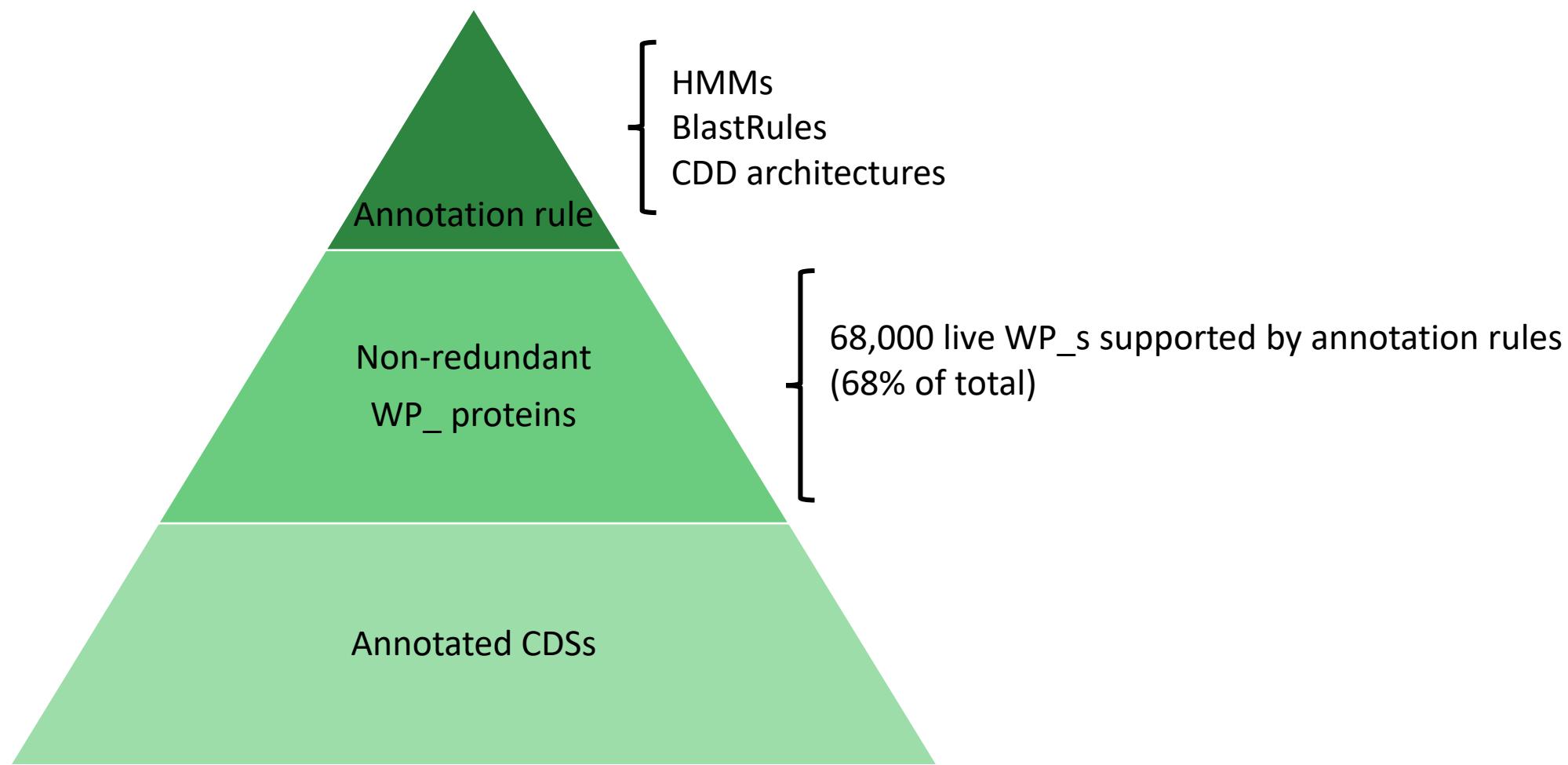
-  bacteria [522]
 -  Escherichia [515]
 -  Shigella [6]
 -  Klebsiella [1]
-  eukaryotes [1]



U.S. National Library of Medicine
National Center for Biotechnology Information



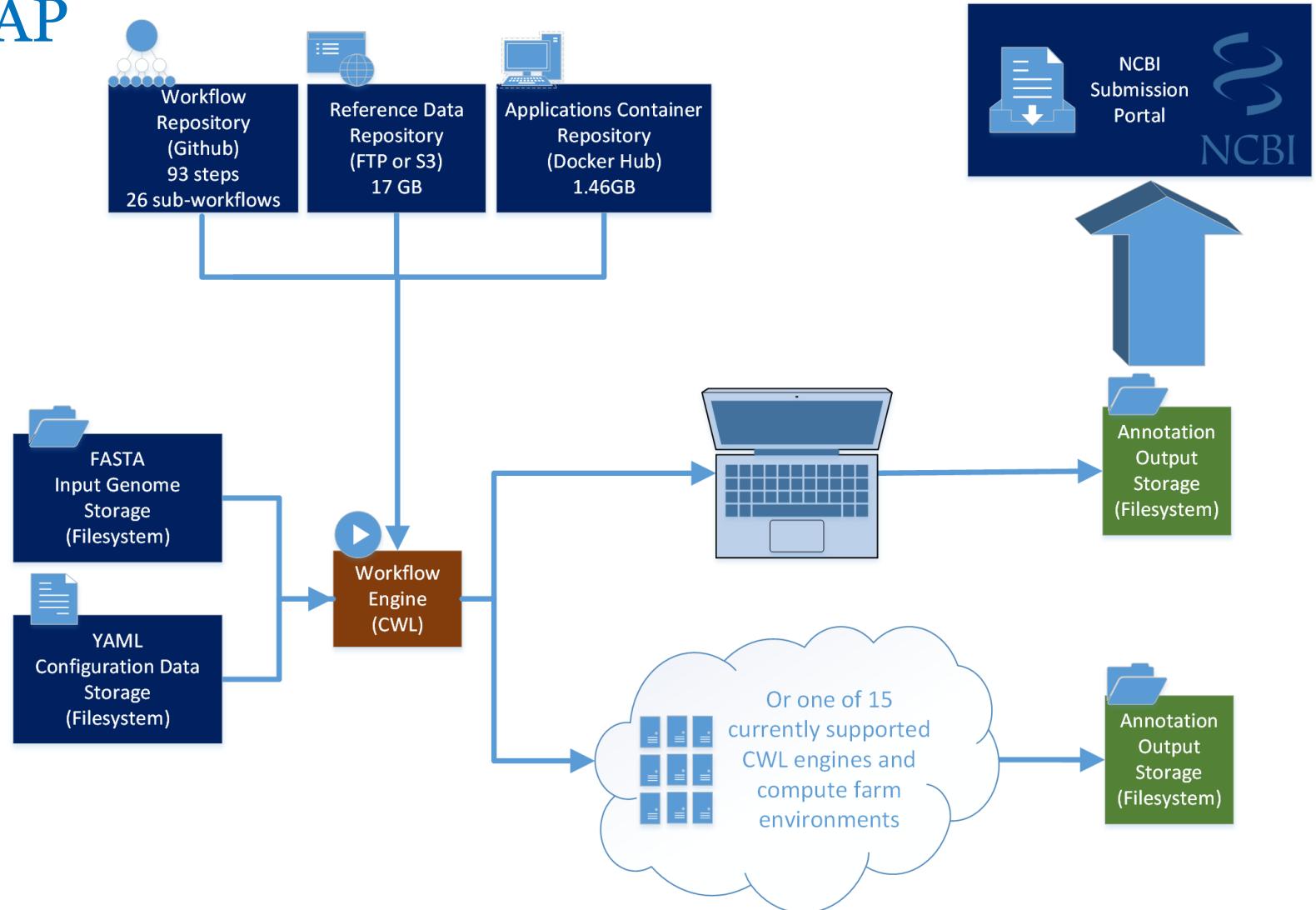
Multiplier effect of annotation rules and non-redundant proteins



Containerization of PGAP

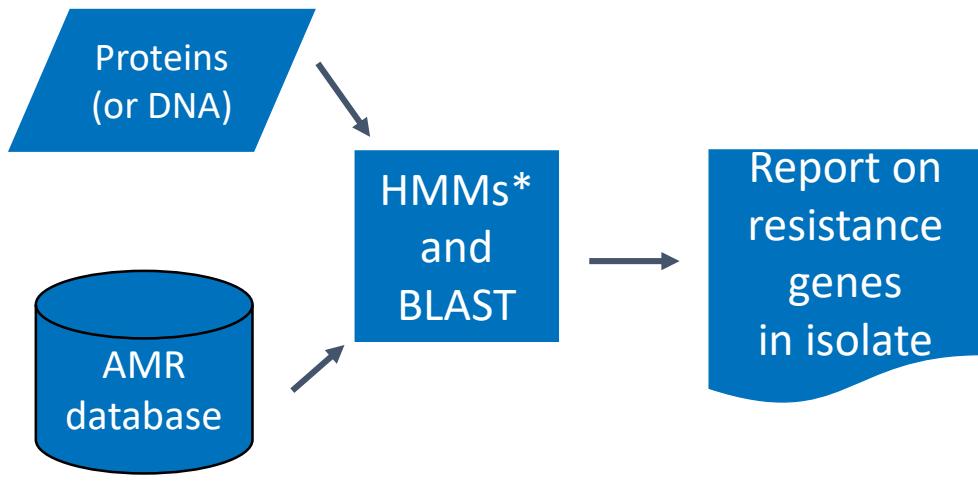
- Conforming with PGAP at NCBI
- No dependencies on NCBI resources
- Reproducible
- Executable by external users on a variety of platforms
- Producing data acceptable to GenBank

<https://github.com/ncbi/pgap>



AMRFinder

Stand-alone tool for the detection of anti-microbial resistance genes



4,730 resistance proteins
565 HMMs
34 drug classes resisted
~50% beta-lactamases

Exact match

Protein name

KPC-2

Functional determination

Resistance to carbapenems and other beta-lactam antibiotics.

HMM score > cutoff of KPC family

KPC family

Likely resistance to carbapenems and other beta-lactam antibiotics.

HMM score > cutoff

class A β -lactamase

Class A beta-lactamase of unknown specificity.

HMM score < cutoff

not beta-lactamase

Prevents false-positive identification as a beta-lactamase. Not reported.

<https://github.com/ncbi/amr/wiki>



U.S. National Library of Medicine
National Center for Biotechnology Information



Conclusions

- Some elements are critical for quality annotation:
 - Quality assemblies
 - Experimental evidence
- Strategies to leverage curated and high-quality data are important
 - Reference annotation
 - Manual and semi-automated curation
- NCBI initiatives to distribute stand-alone tools include the distribution of curated datasets



Thank you.

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

GenBank	GEO	Annotation Pipeline	RefSeq/Gene	GDV/Remap/GBench
Shelby Bidwell	Emily Clough	Francoise Thibaud-Nissen	Eric Cox	Shashi Pujar
Larissa Brown	Carlos Evangelista	Paul Kitts	Catherine Farrell	Bhanu Rajput
Jianli Dai	Irene Kim	Mike Dicuccio	Tamara Goldfarb	Sanjida Rangwala
Scott Durkin	Pierre Ledoux	Wratko Hlavina	Diana Haddad	Lillian Riddick
Michel Eschenbrenner	Hyeseung Lee	Avi Kimchi	John Jackson	Barbara Robbertse
Linda Frisse	Kimberly Marshall		Vinita Joardar	Brian Smith-White
Leigh Riley	Katherine Phillippe	Jinna Choi	Kelly McGarvey	Pooja Strope
	Patti Sherman	Patrick Masterson	Michael Murphy	Anjana Vatsan
BioProject / Biosample	Stephen Wilhite	Eyal Mozes	Nuala O'Leary	David Webb
	Tanya Barrett	Robert Smith	RefSeq Developers	
John Anderson		Alexandre Souvorov	Alex Astashyn	
Carol Scott	GEO developers		Olga Ermolaeva	
	Alexandra Soboleva		Vamsi Kodali	
	Maxim Tomashevsky		Craig Wallin	
	Nadezhda Serova			
	Naigong Zhang			

A cast of thousands

Ken Katz
Michael Ovetsky
Lukas Wagner
Andrei Shkeda
Donna Maglott
Kim Pruitt
Jim Ostell

Watch NCBI News for updates!

<http://www.ncbi.nlm.nih.gov/news/>

<https://www.youtube.com/user/NCBINLM>



U.S. National Library of Medicine
National Center for Biotechnology Information

NCBI Genome Resources Workshop

Monday Jan 14, 12:50-3pm Pacific Salon 2

Time	Topic
12:50 – 1:10	Submission of Genomes to GenBank <i>Karen Clark</i>
1:10 – 1:30	GEO Submissions and Usage <i>Steve Wilhite</i>
1:30 – 1:55	From Annotation to Visualization: Exploring Genes and Genomes with NCBI Tools <i>Eric Cox</i>
1:55 – 2:15	Programmatic Access to Genomic Data: E-Utilities and FTP <i>Vamsi K. Kodali</i>
2:15 – 2:35	NCBI Resources for Phylogenetically-Defined Next Generation Analysis in and out of the Cloud (a.k.a. Cool New Stuff!) <i>Ben Busby</i>
2:35 – 3:00	Q & A session



U.S. National Library of Medicine
National Center for Biotechnology Information

Visit NCBI Booth **223**

Contact us info@ncbi.nlm.nih.gov